Grandes Modelos de Lenguaje: Fundamentos y Aplicaciones

Descripción del Curso:

Este curso explora los conceptos fundamentales, arquitecturas y aplicaciones de los Grandes Modelos de Lenguaje (LLMs, por sus siglas en inglés). Los estudiantes adquirirán un conocimiento profundo de las arquitecturas de los LLM, desarrollarán habilidades en ingeniería de prompts, comprenderán los sistemas de generación aumentada por recuperación (RAG) e investigarán sistemas agenciales que integran los LLM en flujos de trabajo del mundo real.

Requisitos Previos:

- Conocimientos básicos de aprendizaje automático y aprendizaje profundo
- Familiaridad con programación en Python
- Recomendado: Exposición previa al procesamiento de lenguaje natural (NLP)

Objetivos del Curso:

Al final del curso, los estudiantes serán capaces de:

- Comprender la arquitectura y los principios de entrenamiento de los LLM.
- Diseñar prompts efectivos para diversas aplicaciones con LLM.
- Implementar sistemas de generación aumentada por recuperación (RAG).
- Explorar y crear sistemas agenciales que utilicen LLM para la toma de decisiones y la automatización.

Temario:

Módulo 1: Fundamentos de los Grandes Modelos de Lenguaje

- Introducción a los LLM
  - Visión general de los LLM y su evolución
  - Implicaciones sociales y éticas de los LLM

- Arquitectura y Entrenamiento de los LLM
  - Arquitectura Transformer (mecanismo de atención, auto-atención y redes de alimentación hacia adelante)
  - Preentrenamiento, ajuste fino y ajuste por instrucciones
  - Tokenización y embeddings

Módulo 2: Ingeniería de Prompts

- Principios de la Ingeniería de Prompts
    - Diseño efectivo de prompts: zero-shot, one-shot y few-shot
    - Rol del contexto y formato del prompt
    - Práctica: Redacción y refinamiento de prompts

- Ingeniería Avanzada de Prompts
    - Razonamiento en cadena (Chain-of-thought reasoning)
    - Optimización de prompts para tareas específicas de dominio
    - Práctica: Construcción de prompts específicos para tareas

Módulo 3: Generación Aumentada por Recuperación (RAG)

- Introducción al RAG
    - Concepto y beneficios de RAG
    - Componentes clave: recuperador y generador

- Construcción de Sistemas RAG
    - Implementación de bases de datos vectoriales y modelos de embeddings
    - Integración de recuperadores con LLM
    - Práctica: Desarrollo de un sistema RAG simple

- Aplicaciones de RAG
    - Casos de uso: preguntas y respuestas, resumen y recuperación de documentos
    - Desafíos y métricas de evaluación

Módulo 4: Sistemas Agenciales

- Introducción a los Sistemas Agenciales
    - Concepto de comportamiento agencial en sistemas de IA
    - Arquitecturas para sistemas multi-agente

- Diseño de Sistemas Agenciales con LLM

    - Combinación de LLM con herramientas externas (e.g., APIs, bases de datos)
    - Práctica: Creación de agentes específicos para tareas

- Estudios de Caso y Aplicaciones en el Mundo Real
    - Sistemas autónomos en investigación, negocios y atención médica
    - Práctica: Construcción de un flujo de trabajo agencial a pequeña escala

Módulo 5: Temas Avanzados y Direcciones Futuras

- Arquitecturas Avanzadas y Ajuste Fino
    - LoRA y adaptadores
    - Leyes de escalado y optimización de eficiencia

- Implicaciones Éticas y Sociales de los LLM
    - Sesgos, equidad y preocupaciones de seguridad
    - Despliegue responsable de los LLM

Evaluación y Calificaciones:

- **Tareas (40%)**: Ejercicios semanales de programación sobre ingeniería de prompts, RAG y sistemas agenciales.
- **Proyecto Intermedio (20%)**: Diseñar e implementar una aplicación aumentada por recuperación.
- **Proyecto Final (30%)**: Desarrollar un sistema agencial funcional que resuelva un problema del mundo real.
- **Participación (10%)**: Contribución en discusiones de clase y actividades colaborativas.

Recursos y Herramientas:

- **Frameworks de Programación:** Python, PyTorch/TensorFlow, Hugging Face Transformers
- **Bases de Datos:** Pinecone, Weaviate o FAISS
- **Modelos Preentrenados:** OpenAI, GPT, T5, LLaMA, etc.
- **Lecturas:** Artículos y documentos seleccionados (e.g., “Attention is All You Need,” artículos sobre RAG, etc.)